When Biological Evidence Misleads: Reliability-Aware Transfer for Cross-Scale Drug Response Prediction

Anonymous authors
Paper under double-blind review

Abstract

External biological evidence can enhance drug-response prediction, but its availability does not guarantee that transferring it will benefit unseen compounds. We investigate selective evidence transfer, in which an auxiliary biological-evidence channel is admitted or suppressed according to its compatibility with the training distribution. Our framework distinguishes the effects of evidence-aware training from the inference-time contribution of evidence within a fixed model. Using frozen drug–protein representations and a pathway-coverage-based admission rule, we evaluate transfer across 22 drugs, including eight with mapped biological evidence, under held-out-drug and scaffold-disjoint protocols. Matched three-seed comparisons show that the support-based rule reduces the largest observed drug-specific performance losses relative to conventional fusion approaches. However, an eight-seed extension reveals uncertain mean improvement and increased sensitivity of the worst-drug outcome to training randomness. Inference-time corruption further demonstrates that partial evidence removal can reactivate transfer for an initially rejected drug, while gate magnitude does not exhibit a statistically resolved association with drug-level utility. These findings establish the importance of distinguishing evidence availability, training-level effects, and inference-time utility when evaluating selective transfer. Our framework provides an auditable approach to identifying negative-transfer risks while highlighting the limitations of coverage-based reliability decisions under drug-distribution shift.

1 Introduction

External biological knowledge is increasingly incorporated into drug-response prediction through drug–target interactions, pathway annotations, and pretrained molecular representations. These sources offer a way to connect molecular binding evidence to cell-line response when target-task labels are limited. Their usefulness is nevertheless conditional: an interaction measured in a binding assay need not provide a transferable signal in a particular cellular context. A predictor can therefore benefit from an auxiliary source for one compound while losing accuracy for another. An aggregate improvement alone does not reveal whether this trade-off is acceptable or which part of the transfer process creates it.

This problem becomes more pronounced when the drugs used for evaluation differ from those with response labels during training. Evidence coverage varies across compounds, and densely studied drugs may have support patterns that are unusual relative to a small training cohort. A learned gate can respond strongly to such evidence without estimating whether it is useful. Conversely, a model may appear robust to evidence corruption simply because it uses very little of the auxiliary signal. These possibilities motivate an evaluation that makes evidence use, realized utility, and the consequences of suppression separately observable.

We formulate *selective evidence transfer*: the response predictor always returns a prediction, while a reliability decision determines whether a biological-evidence channel contributes. Our implementation freezes a BindingDB-pretrained drug–protein representation and aggregates it into thirteen pathway slots. A deliberately simple support rule suppresses evidence when its number of pop-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

ulated slots exceeds the maximum among supported training drugs. This rule is an interpretable reference mechanism, not a general estimator of evidence correctness or predictive benefit. It lets us examine what transfer control can gain and where it still fails when coverage is only a limited proxy for reliability.

We draw a central distinction between two comparisons that are often conflated under a single no-evidence counterfactual. Independently training a model without evidence measures the net effect of an evidence-aware training procedure, including changes to shared parameters. Turning evidence off in a fixed checkpoint measures the channel's inference-time contribution while holding those parameters constant. The two quantities need not agree. In particular, suppressing an auxiliary update does not undo the effect that evidence has already had on backbone training. This distinction is necessary for interpreting fallback behavior and for deciding whether a mechanism truly prevents negative transfer.

Our contributions are threefold. First, we formulate selective evidence transfer as a decision about auxiliary information rather than prediction rejection, with separate training-level and fixed-checkpoint reference comparisons. Second, we introduce a training-support-based instantiation and characterize the scope of its channel-suppression property, contrasting joint training with an isolated residual that retains a frozen base. Third, we establish a drug-level reliability evaluation combining chemical-shift protocols, seed-matched controls, harm counts, worst-drug sensitivity, and corruption interventions. Together, these components identify when transfer control reduces observed drug-specific losses and when coverage-based suppression remains insufficient.

2 Related Work

Transfer learning provides a framework for adapting information across domains with different data availability and distributions (Pan & Yang, 2010). In drug-response prediction, few-shot transfer, CODE-AE, pathway-based transfer, and scDEAL address complementary aspects of adaptation from experimental systems to target cohorts (Ma et al., 2021; He et al., 2022; Tang et al., 2022; Chen et al., 2022). TransCDR and knowledge-guided domain adaptation incorporate multimodal or structured information (Xia et al., 2024; Liu & Li, 2025), while DrugCell, GPDRP, and condition-specific interaction pretraining investigate biological priors for response prediction (Kuenzi et al., 2020; Yang & Li, 2023; Bang et al., 2024). DeepSADR provides the interaction and readout design underlying the response backbone (Zhang et al., 2026); the adaptive-readout control evaluated here is a protocol-matched variant rather than a reproduction of the complete published system. Building on these representation and adaptation approaches, we examine when external evidence should contribute to prediction under drug-distribution shift.

Negative transfer characterizes a loss relative to target-only learning, and existing methods formalize this comparison and reduce the influence of unrelated source information (Wang et al., 2019). Task-uncertainty-guided adaptation addresses related limitations in cancer drug-response transfer (Peres da Silva et al., 2021). Selective prediction controls a risk–coverage trade-off by rejecting predictions (Geifman & El-Yaniv, 2019; Fisch et al., 2022; Liang et al., 2024), whereas learning to defer assigns selected decisions to an expert (Mozannar & Sontag, 2020). Selective evidence transfer differs in the object of selection: the auxiliary representation can be suppressed while the response predictor continues to produce an output. Evaluating this intervention requires both a separately trained target-only reference, which captures the net effect of evidence-aware training, and an evidence-off computation of the same checkpoint, which isolates the channel's inference-time contribution.

Reliability of auxiliary information also connects OOD detection with robust multimodal learning. Feature-based detectors identify departures from training distributions through scores such as Mahalanobis confidence (Lee et al., 2018), but unfamiliarity need not imply negative predictive utility. MultiBench evaluates missing and noisy inputs (Liang et al., 2021); modality competition and quality-aware dynamic fusion address optimization and input-quality limitations (Huang et al., 2022; Zhang et al., 2023). These approaches motivate an explicit assessment of auxiliary reliability. We examine drug-level negative transfer through two reference computations and evidence perturbations, including coverage-preserving controls. Our scalar support rule provides an interpretable transfer-selection mechanism, not a general OOD detector.

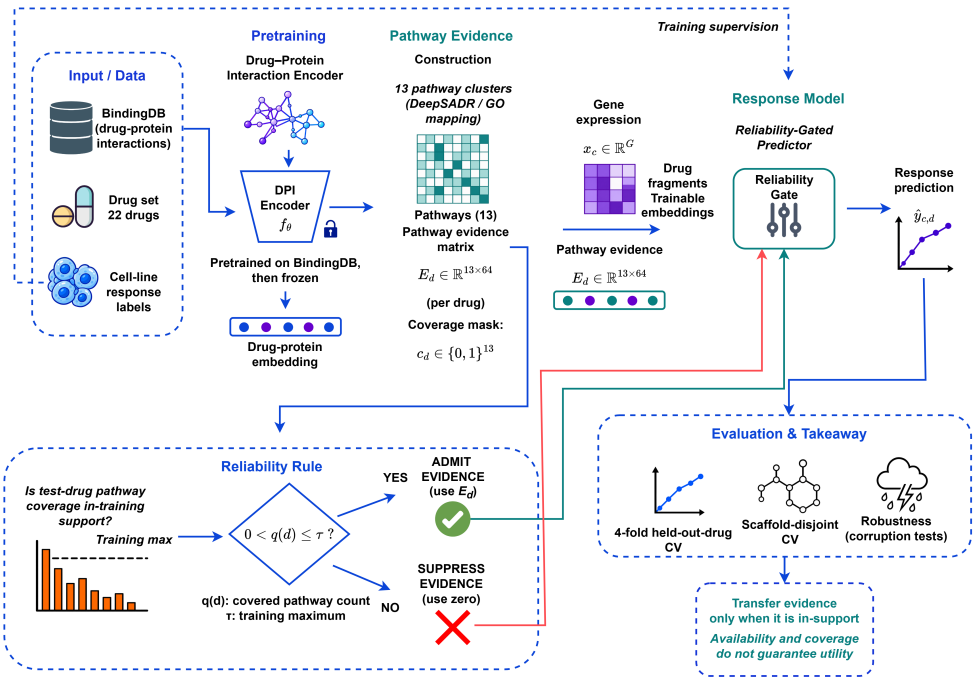


Figure 1: Selective evidence-transfer pipeline. A frozen binding-pretrained encoder supplies pathway evidence, admitted or suppressed by a training-coverage rule. Gene expression and trainable drug-fragment embeddings feed the response model, whose gated fusion precedes an interaction graph, masked-mean readout, and prediction head (details omitted in the schematic). Suppression retains the same checkpoint; it does not recover an independently trained no-evidence model.

3 METHOD

3.1 EVIDENCE CONSTRUCTION AND RESPONSE REPRESENTATION

Let x contain a cell-line expression profile and a drug representation, and let d denote the drug. For pathway $k \in \{1, \dots, 13\}$, $T_{d,k}$ is the set of retained positive BindingDB target matches after structural and pathway mapping. Given frozen pair embedding v_j and weight $w_j = \mathrm{sim}_j \mathrm{obs}_j$, the pathway evidence and coverage are defined as

$$e_{d,k} = \frac{\sum_{j \in T_{d,k}} w_j v_j}{\sum_{j \in T_{d,k}} w_j}, \qquad c_{d,k} = \mathbb{I}[|T_{d,k}| > 0],$$

setting $e_{d,k} = 0$ when no match is available. Frozen MolFormer and ESM-2 embeddings feed a pair encoder trained on BindingDB IC50 supervision (Ross et al., 2022; Lin et al., 2023; Gilson et al., 2016). The trained pair encoder and resulting evidence tensors are frozen before response-model fitting. The response backbone separately uses trainable drug-fragment embeddings and an expression-derived pathway encoder; the frozen molecular encoders belong to the evidence branch, separate from this trainable fragment lookup.

Let s_k be the sample-specific pathway feature and $\bar{d}$ the pooled response-side drug context. The implemented fusion operates *per pathway*, before the interaction graph:

$$a_k = \sigma\big(g([s_k; e_{d,k}; \bar{d}; |s_k - e_{d,k}|; c_{d,k}])\big)\, c_{d,k}, \qquad h_k = s_k + \beta a_k W_e e_{d,k}.$$

Here g is a two-layer MLP, W_e is bias-free, and the trainable scalar β starts at zero. The graph, readout, and prediction head operate on these updated pathway features and the drug features. Joint optimization can consequently change the downstream predictor even for samples whose evidence is later suppressed. Zero initialization removes the initial auxiliary update; it does not initialize from a previously trained target-only checkpoint.

3.2 SELECTIVE SUPPORT DECISION AND FALLBACK BOUNDARIES

Define density $q(d) = \sum_k c_{d,k}$ and let $\mathcal{S}_{\text{train}}$ be the supported drugs present in a response-training fold. The support threshold and decision are

$$\tau = \max_{d' \in \mathcal{S}_{\text{train}}} q(d'), \qquad A(d) = \mathbb{I}[0 < q(d) \leq \tau].$$

The term OOD abstention denotes this one-sided coverage-support rule: $A(d) = 1$ admits evidence, whereas $q(d) > \tau$ triggers suppression. The rule is not a general distribution-support test; it does not reject positive coverage below the training minimum or gaps within the observed range. For $q(d) > \tau$, the implementation zeros the evidence and coverage before fusion, making the applied gate zero. For $q(d) = 0$, the coverage mask already prevents transfer. Both the learned-gate control and OOD abstention use no density normalization; the key difference between them is the hard support decision. The threshold is derived from training support and held fixed at test time, with no test response labels used. Although the decision is evaluated during training as well, no training-supported drug exceeds its own fold maximum.

For finite inputs in evaluation mode, disabling the channel gives $h_k = s_k$. It follows that the prediction equals the *evidence-off computation of that same checkpoint*. This is an implementation identity, not a statistical guarantee about accuracy or equality to another trained model. The maximum is the smallest upper threshold that leaves every supported training drug admissible, but this elementary property does not establish an optimal threshold for transfer utility. Accordingly, the rule serves as an interpretable reference mechanism within the reliability-audit framework.

Provenance weighting is a secondary analysis: for an admitted drug it replaces binary coverage with $\min(1, n_{d,k}/\kappa)$, where $n_{d,k}$ counts retained targets and κ is the training-fold 75th percentile over covered pathway counts. The support check still uses binary density. This adds support-strength information but does not itself calibrate predictive utility. The isolated-residual control provides a different guarantee. It first trains a no-evidence base, freezes its parameters, and then learns an additive *logit* residual on supported training rows:

$$f_{\text{iso}}(x, e) = f_{\text{base}}(x) + A(d)\rho_\phi(x, e).$$

If $A(d) = 0$, its logit is exactly that of its own frozen base. This guarantee follows from retaining the base parameters and masking the residual, and does not imply positive utility on admitted samples. It does not transfer to the jointly trained primary method.

3.3 TRAINING AND INFERENCE PROCEDURE

Algorithm 1 separates frozen evidence from trainable response components. For the primary method, the expression encoder, fragment embeddings, interaction backbone, prediction head, and gate are jointly optimized with response BCE and a KL regularizer.

Each method is fitted separately with the same fold manifest and named seeds. The no-evidence comparison supplies zero evidence throughout its own training; its weights are not copied into OOD abstention. A separate no-evidence control, adaptive readout, supplies zero evidence but replaces the default masked-mean pathway readout with a learned attention-based one, isolating whether the readout architecture rather than the evidence channel explains any part of the no-evidence baseline. Static addition, concatenation, and cross-attention share backbone dimensions and the response-head design, but their fusion modules have different parameter counts (Table 3): static addition adds none, concatenation and cross-attention add roughly sixteen to seventeen thousand, and the gate shared by OOD abstention, the unnormalized-gate control, provenance weighting, and the SelectiveNet-style objective adds about twenty-one thousand. They are architecture controls, not exactly parameter-matched models.

Corruption is applied to frozen checkpoints without refitting either weights or thresholds. Removal deletes selected supported slots and their coverage. Pathway shuffling cyclically reassigns a selected subset of occupied slots while preserving the coverage count; a drug with only one occupied slot cannot undergo a nontrivial such permutation. Noise mixes evidence with matched-scale Gaussian values, whereas drug shuffling assigns another drug's evidence tensor. These controls perturb different properties, so their nominal severity fractions should not be interpreted as equivalent biological damage.

Algorithm 1 Training and inference for selective evidence transfer.

Require: binding data, response folds, seed list, drug–pathway mapping

Cache frozen MolFormer/ESM-2 embeddings; train the pair encoder on binding supervision.
Freeze the pair encoder; construct (e_d, c_d) for every response drug.
for each response fold and seed **do**
  Compute τ from supported training drugs only.
  Initialize and train an independent no-evidence baseline using zero tensors.
  Initialize response parameters θ, gate ϕ, and $\beta = 0$ for transfer.
  for each training minibatch **do**
    Encode response features; apply $A(d)$; form h_k; jointly optimize (θ, ϕ, β) using BCE plus KL.
  end for
  Save checkpoint, threshold, test-row membership, and predictions.
end for
Inference: fix all weights and τ; recompute $A(d)$ from supplied evidence. If $A(d) = 0$, suppress the auxiliary update; always return a response prediction.
Audit: compare independently trained models on matched seeds/rows; separately hold the transfer checkpoint fixed and corrupt or zero its evidence.
Optional isolated residual: freeze a fitted base before training ρ_ϕ; add its logit only when $A(d) = 1$.

3.4 Two reference comparisons and drug-level reliability

Let f_{θ_e} be an evidence-trained checkpoint and f_{θ_0} a separately fitted no-evidence model. Training-level utility and within-checkpoint utility are defined separately:

$$U_{\text{train}}(d) = \text{AUC}_d[f_{\theta_e}(x, e)] - \text{AUC}_d[f_{\theta_0}(x, 0)],$$
$$U_{\text{off}}(d) = \text{AUC}_d[f_{\theta_e}(x, e)] - \text{AUC}_d[f_{\theta_e}(x, 0)].$$

The first includes effects of training with evidence on shared weights; the second isolates enabling the channel after training. Matched rows and seeds reduce comparison variability but do not make the optimization trajectories or final weights identical. Consequently, U_{train} and U_{off} may differ, and zero applied gate need not mean zero training-level utility. Neither comparison establishes a causal biological effect of a target interaction.

For supported drugs $\mathcal{S}$, the summaries of $U = U_{\text{train}}$ comprise its mean, median, worst value $W = \min_{d \in \mathcal{S}} U(d)$, and harm count $H_\delta = \sum_{d \in \mathcal{S}} \mathbb{I}[U(d) < -\delta]$. We term the signed minimum worst utility; conventional nonnegative regret would instead be $\max(0, -U)$. Thresholds $\delta \in \{0, .01, .02, .05\}$ are exploratory absolute-AUC cutoffs, not clinically validated margins. This grid distinguishes any decrease from larger decreases without choosing a cutoff that favors one method.

Paired seed deltas are averaged within each fold before computing fold-bootstrap summaries of mean utility. Because cross-validation training sets overlap and there are only four folds, these intervals are descriptive and do not establish population-level significance. For worst-drug sensitivity, paired training seeds are resampled within each fixed fold and the minimum is recomputed over all eight drugs. These conditional intervals describe training-randomness sensitivity on the observed cohort, not uncertainty for a new drug population or new patient rows. Leave-one-drug-out (LODO) ranges recompute the minimum after omitting each drug, without refitting, to show how much a single compound determines the lower tail.

4 Results

4.1 Data and evaluation protocols

The primary cell-line benchmark contains 13,282 response rows across 22 drugs, eight with positive mapped evidence at the conservative structural-matching threshold. The patient resource contains 338 rows across five drugs, but a provenance audit resolves only 186 rows to the TCGA cohort; the

Table 1: Held-out-drug comparison, four folds and three common seeds. U compares separately trained models; $H_{.01}$ counts supported drugs losing more than .01 AUC. Intervals and other cutoffs appear in Appendix B.

Method	AUC	Mean U	Median U	Worst U	$H_{.01}$
No evidence	0.5889	+0.0000	+0.0000	+0.0000	0/8
Adaptive readout (zero)	0.5866	-0.0023	+0.0023	-0.0864	3/8
Static addition	0.5707	-0.0182	-0.0002	-0.1588	2/8
Concatenation	0.5595	-0.0294	-0.0044	-0.2520	2/8
Cross-attention	0.5838	-0.0051	-0.0032	-0.0349	2/8
Learned gate	0.5836	-0.0053	-0.0002	-0.0756	1/8
OOD abstention	0.5956	+0.0066	+0.0035	-0.0016	0/8
Provenance	0.5949	+0.0059	+0.0003	-0.0043	0/8
SelectiveNet-style	0.5701	-0.0189	+0.0018	-0.1855	1/8

Table 2: Scaffold-disjoint comparison on the same three seeds for every method. The eight-seed no-evidence/OOD pair is reported separately in the text.

Method	AUC	Mean U	Median U	Worst U	$H_{.01}$
No evidence	0.5959	+0.0000	+0.0000	+0.0000	0/8
Static addition	0.5856	-0.0104	-0.0009	-0.0837	1/8
Concatenation	0.5835	-0.0124	+0.0009	-0.1296	1/8
Cross-attention	0.5914	-0.0045	-0.0006	-0.0247	1/8
Learned gate	0.5735	-0.0224	+0.0037	-0.2099	1/8
OOD abstention	0.5997	+0.0038	+0.0037	-0.0021	0/8
Provenance	0.5965	+0.0006	+0.0003	-0.0038	0/8

other 152 pseudonymous rows do not support verified cancer metadata. Patient evidence is concentrated in sorafenib. The cell-line benchmark therefore provides the primary evidence for the reliability analysis; patient transfer and cancer-specific generalization remain unestablished. Biological-expression and binding resources are described in Ghandi et al. (2019); Gilson et al. (2016); cohort construction and exclusions are documented in Appendix A.

Held-out-drug evaluation uses four drug-disjoint folds, with each drug tested once and two supported drugs in each test fold. Scaffold-disjoint evaluation groups compounds by Bemis–Murcko scaffold (Bemis & Murcko, 1996). Disjoint scaffold identifiers constrain one source of chemical overlap but do not prove that all test compounds are distant from training compounds. Every cross-method table below uses the common seeds 50–52 in every fold. The separate eight-seed extension uses seeds 50–57 for no evidence and OOD abstention only. Seed alignment precedes averaging in each comparison, ensuring that differences in seed composition do not confound cross-method estimates.

The main inference-only corruption suite comprises four held-out-drug folds and three seeds per fold, with 15 conditions per trained checkpoint. A legacy five-seed patient corruption experiment used an earlier representation and exhibited nearly collapsed predictions. It is retained in the development record and excluded from the current mechanism evidence. The response model uses 64-dimensional features, masked-mean readout, AdamW, ten epochs, and BCE plus a KL penalty; implementation details are in Appendix A. No additional training was performed to obtain the matched-seed and harm-sensitivity summaries.

4.2 Matched comparisons and sensitivity of drug-level harm

Table 1 reports the held-out-drug results. OOD abstention has supported-drug macro-AUC 0.5956 versus 0.5889 for no evidence, with mean utility $+0.0066$ and a fold-bootstrap interval approximately $[-0.0007, +0.0140]$. This interval includes zero. The drug-level contrast is more pronounced: static addition and concatenation have worst utilities -0.1588 and -0.2520, whereas OOD abstention has -0.0016. Nevertheless, three of the eight drugs still have negative mean deltas under OOD abstention. Its zero count at a .01-AUC cutoff describes the magnitude of these observed losses, not elimination of negative transfer.

The matched scaffold comparison in Table 2 gives OOD abstention mean utility $+0.0038$ and worst utility -0.0021, with two drugs harmed and none losing more than .01 AUC. The learned gate has a positive median but a negative mean, driven by a worst utility of -0.2099. This illustrates why a median, an average, and a harm count answer different questions. Provenance has mean utility $+0.0006$ and worst utility -0.0038 on these same seeds. On the matched three-seed comparison, provenance has a smaller mean benefit and a more negative minimum than OOD abstention; it remains a secondary reliability analysis.

In the eight-seed extension, OOD abstention reaches 0.5959 versus baseline 0.5951, with mean utility $+0.0009$ and fold-bootstrap interval $[-0.0011, +0.0036]$. Worst utility becomes -0.0111 and $H_{.01} = 1/8$. Although the three-seed mean interval excludes zero, the eight-seed interval includes zero. This seed sensitivity limits the evidence for a consistent mean improvement. Across all drugs, the eight-seed overall macro-AUC delta is approximately -0.0023, further limiting a performance-based interpretation.

Conditional paired-seed intervals for worst utility are approximately $[-0.0037, -0.0009]$ for held-out-drug OOD abstention, $[-0.0037, -0.0013]$ for its three-seed scaffold result, and $[-0.0271, -0.0028]$ for its eight-seed extension. The wider latter interval shows that near-zero observed harm in a small seed sweep need not be stable. The corresponding scaffold LODO range changes from $[-0.0021, -0.0001]$ at three seeds to $[-0.0111, -0.0049]$ at eight seeds. Full method-level intervals, harm counts at all four cutoffs, and per-drug values are provided in the appendix. These are sensitivity diagnostics conditional on the current cohort, not guarantees about future compounds.

4.3 Corruption, gate utility, and structural controls

The fixed-checkpoint audit holds the learned parameters and support threshold constant while changing evidence at inference. Observed supported AUC is 0.5956, whereas zero evidence within the same checkpoints gives 0.5915. This differs from the independently trained no-evidence AUC of 0.5889, empirically illustrating the two comparisons in Section 3. Complete removal zeros the channel by construction; it does not reverse the training process. The partial-removal curve is non-monotone: removing 25% of supported slots gives AUC 0.5835, while complete removal gives 0.5915.

Sorafenib provides the clearest failure mechanism. Its original coverage is outside training support, so the rule suppresses its evidence. Removing 25% lowers coverage *into* the accepted range, re-activates transfer with mean applied gate 0.4133, and reduces its AUC from 0.7269 to 0.6529. At complete removal the coverage mask is zero and the AUC returns to 0.7269. Thus, a support check can reject unusually dense evidence yet accept a corrupted subset of it. This result explains a specific missingness failure; it does not establish that all partial corruption is more harmful than complete absence.

Pathway shuffling changes supported macro-AUC by at most approximately .0002 over the tested severities; full noise and drug shuffle give deltas around $-.0029$ and $-.0037$. The small pathway effect is consistent with weak sensitivity to slot assignment, but does not prove biological invariance: the operator permutes occupied slots, some drugs have few movable slots, and aggregation may reduce sensitivity. These controls cannot exclude every coverage confound or establish that pathway identity is irrelevant in another model. The corruption table in Appendix D keeps the operator and severity explicit.

Figure 2 directly compares mean applied gate and training-level utility across eight supported drugs. Held-out-drug Spearman correlation is -0.143 ($p = .736$); the eight-seed scaffold correlation is .333 ($p = .420$). The analyzed scores are the applied OOD-abstention gates, including the support mask, rather than unconstrained gate outputs. A rank association is also not a calibration test. The current evidence provides no reliable basis for using gate magnitude as an estimate of benefit, but the small drug count cannot rule out an association in a larger cohort. Correlation with training-level utility additionally includes backbone changes that an inference-time gate cannot undo.

The isolated residual yields exact equality to its frozen base on inactive samples, verified in all twelve scaffold runs, but its mean supported utility is -0.0003 with interval $[-0.0063, +0.0045]$. It demonstrates a structural fallback property without establishing a benefit for admitted drugs. A

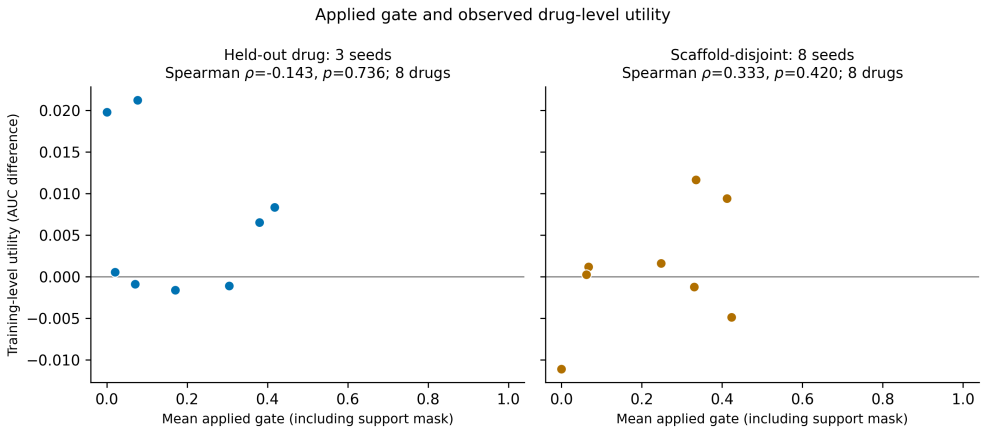


Figure 2: Applied gate magnitude versus training-level drug utility for OOD abstention. Held-out drug uses three seeds; the scaffold panel uses the eight-seed extension. Each point is one supported drug. Neither correlation is statistically resolved; these eight-point analyses do not prove calibration or its impossibility.

SelectiveNet-style gate-coverage penalty gives held-out supported AUC 0.5701 and worst utility -0.1855. This implementation encourages mean gate coverage and lacks the train-max support rule; it is not the complete selective-classification method and does not show that a penalty overrides an existing hard abstention mask. Together, these controls distinguish parameter isolation, encouraged gate use, and hard channel suppression as different interventions.

5 Discussion

Our central contribution is an auditable formulation of selective evidence use. An evidence channel can be available, familiar under a chosen support statistic, and still unhelpful for response prediction. Treating these as separate properties leads to concrete evaluation obligations: specify the baseline training procedure, measure the fixed-checkpoint contribution, report which drugs are harmed, and challenge the decision with corruptions that preserve nominal support. The scalar threshold is useful because its behavior is inspectable. Its simplicity also makes the failure under partial removal interpretable, without suggesting that pathway count is a sufficient representation of reliability.

The distinction between the two reference computations changes what can be claimed about fallback. Suppression ensures that an auxiliary update vanishes in the current forward pass. Only freezing the base before fitting an isolated residual ensures that inactive predictions preserve that particular base. Evidence-aware joint training can otherwise change shared parameters and affect unsupported drugs as well. The unfavorable overall macro-AUC delta in the eight-seed scaffold extension shows why lower observed harm on supported drugs must be considered together with this training-level effect. Neither a small minimum loss nor a closed gate constitutes a global no-harm guarantee.

Worst-drug results require particular restraint. A minimum over eight compounds is sensitive to drug composition and training randomness, and one extreme drug can dominate comparisons. The harm-count grid and LODO ranges expose that dependence, while the larger seed sweep reduces the estimated mean benefit. Increasing seeds cannot replace expanding the drug cohort; resampling seeds cannot estimate performance on new chemistry. Similarly, resampling four folds with overlapping training sets supplies a descriptive uncertainty summary rather than four fully independent replications of the research hypothesis. The sensitivity analysis therefore strengthens transparency more directly than it strengthens a claim of universal superiority.

Several limitations concern the evidence source itself. Positive target matches reflect assay availability and a conservative mapping threshold, not a random sample of drug biology. Pathway membership is a coarse abstraction of cellular context, and the current strict construction leaves fourteen drugs unsupported. Sparse patient support and incomplete provenance prevent broad claims about

transfer to clinical outcomes. Finally, prior quality-aware fusion and OOD methods offer richer reliability signals than the present count heuristic; they are relevant future comparators, not methods shown inferior by these experiments. A useful next step is to define a training-side utility target for a reliability estimator and test it on a larger supported cohort while keeping the two reference comparisons explicit.

6 Conclusion

We formulate selective biological-evidence transfer and evaluate it through separate training-level and fixed-checkpoint comparisons. A simple training-support rule reduces the largest observed drug-specific losses in matched three-seed experiments, while additional seeds and partial-corruption tests reveal its limitations. The resulting contribution is a reliability-audit framework with an interpretable reference mechanism and explicit fallback boundaries. The experiments do not establish calibrated gate utility, consistent mean improvement, or clinical effectiveness; they identify what must be measured before external evidence can be treated as reliably transferable.

7 Ethics statement

This study uses existing biological and response resources and makes no treatment recommendation or clinical-validity claim. Patient provenance limitations and negative results are disclosed; harm refers to predictive AUC loss, not measured clinical harm.

8 Reproducibility statement

We document the datasets, experimental protocols, model configurations, evaluation procedures, and statistical analyses to facilitate reproducibility. The appendix provides implementation details, decision thresholds, interval definitions, and protocol-specific random seeds. Reported statistical analyses were performed using stored experimental outputs without retraining. Cross-method consistency was assessed by verifying drug membership, supported-drug indicators, per-drug sample counts, and the non-overlapping partition of all 22 drugs across the four held-out-drug folds. All reported comparisons passed these checks. However, these checks do not establish row-level prediction identity. Source code, analysis scripts, and trained model checkpoints are not publicly released; therefore, independent reproduction of the complete experimental pipeline requires reimplementation based on the documented methodology.

9 AI use statement

LLM-based assistants, including Claude and OpenAI Codex, assisted with manuscript drafting and editing, literature discovery, code development and debugging, analysis scripts, and interpretation of saved experimental outputs. AI assistance supported plotting code for statistical figures. The pipeline schematic was drawn by the author in draw.io, with AI-assisted feedback on its labels and scientific consistency. The numerical results are computed from experimental artifacts rather than generated as illustrative data. Verification includes matched-seed analysis, checks of drug membership and per-drug row counts, and recomputation of parameter counts and gate–utility correlations; these checks do not establish row-level prediction identity or independently validate all source data. The authors remain responsible for the scientific claims, references, code, and final disclosure of AI use.

REFERENCES

Dongmin Bang, Bonil Koo, and Sun Kim. Transfer learning of condition-specific perturbation in gene interactions improves drug response prediction. *Bioinformatics*, 40(Suppl_1):i130–i139, 2024. doi: 10.1093/bioinformatics/btae249. URL https://pubmed.ncbi.nlm.nih.gov/38940127/.

Guy W. Bemis and Mark A. Murcko. The properties of known drugs. 1. molecular frameworks. *Journal of Medicinal Chemistry*, 39(15):2887–2893, 1996. doi: 10.1021/jm9602928. URL https://pubs.acs.org/doi/10.1021/jm9602928.

Junyi Chen, Xiaoying Wang, et al. Deep transfer learning of cancer drug responses by integrating bulk and single-cell rna-seq data. *Nature Communications*, 13:6494, 2022. doi: 10.1038/s41467-022-34277-7. URL https://www.nature.com/articles/s41467-022-34277-7.

Adam Fisch, Tommi S. Jaakkola, and Regina Barzilay. Calibrated selective classification. *Transactions on Machine Learning Research*, 2022. URL https://arxiv.org/abs/2208.12084.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 2151–2159, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Mahmoud Ghandi, Franklin W. Huang, Judit Jané-Valbuena, Gregory V. Kryukov, Christopher C. Lo, E. Robert McDonald, Jordi Barretina, Ellen T. Gelfand, Craig M. Bielski, Haoxin Li, et al. Next-generation characterization of the cancer cell line encyclopedia. *Nature*, 569(7757):503–508, 2019. doi: 10.1038/s41586-019-1186-3. URL https://www.nature.com/articles/s41586-019-1186-3.

Michael K. Gilson, Tiqing Liu, Michael Baitaluk, George Nicola, Linda Hwang, and Jenny Chong. Bindingdb in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Research*, 44(D1):D1045–D1053, 2016. doi: 10.1093/nar/gkv1072. URL https://academic.oup.com/nar/article/44/D1/D1045/2502601.

Di He, Qiao Liu, You Wu, and Lei Xie. A context-aware deconfounding autoencoder for robust prediction of personalized clinical drug response from cell-line compound screening. *Nature Machine Intelligence*, 4(10):879–892, 2022. doi: 10.1038/s42256-022-00541-0. URL https://www.nature.com/articles/s42256-022-00541-0.

Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. Modality competition: What makes joint training of multi-modal network fail in deep learning? (provably). In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 9226–9259, 2022. URL https://proceedings.mlr.press/v162/huang22e.html.

Brent M. Kuenzi, Jisoo Park, Samson H. Fong, Kyle S. Sanchez, John Lee, Jason F. Kreisberg, Jianzhu Ma, and Trey Ideker. Predicting drug response and synergy using a deep learning model of human cancer cells. *Cancer Cell*, 38(5):672–684.e6, 2020. doi: 10.1016/j.ccell.2020.09.014. URL https://pubmed.ncbi.nlm.nih.gov/33096023/.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/abdeb6f575ac5c6676b747bca8d09cc2-Abstract.html.

Hengyue Liang, Le Peng, and Ju Sun. Selective classification under distribution shifts. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2405.05160.

Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Chen, Peter Wu, Michelle A. Lee, Yuke Zhu, Ruslan Salakhutdinov, and Louis-Philippe Morency. MultiBench: Multiscale benchmarks for multimodal representation learning. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://

datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/hash/37693cfc748049e45d87b8c7d8b9aacd-Abstract-round1.html.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/10.1126/science.ade2574.

Xuan Liu and Menglu Li. Knowledge-guided domain adaptation model for transferring drug response prediction from cell lines to patients. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(1):523–531, 2025. URL https://dl.acm.org/doi/10.1609/aaai.v39i1.32032.

Jianzhu Ma, Samson H. Fong, Yunan Luo, Christopher J. Bakkenist, John Paul Shen, Soufiane Mourragui, Lodewyk F. A. Wessels, Marc Hafner, Roded Sharan, Jian Peng, and Trey Ideker. Few-shot learning creates predictive models of drug response that translate from high-throughput screens to individual patients. *Nature Cancer*, 2(2):233–244, 2021. doi: 10.1038/s43018-020-00169-2. URL https://pubmed.ncbi.nlm.nih.gov/34223192/.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 7076–7087, 2020. URL https://proceedings.mlr.press/v119/mozannar20b.html.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010. doi: 10.1109/TKDE.2009.191. URL https://ieeexplore.ieee.org/document/5288526.

Rafael Peres da Silva, Chayaporn Suphavilai, and Niranjan Nagarajan. Tugda: task uncertainty guided domain adaptation for robust generalization of cancer drug response prediction from in vitro to in vivo settings. *Bioinformatics*, 37(Suppl_1):i76–i83, 2021. doi: 10.1093/bioinformatics/btab299. URL https://pubmed.ncbi.nlm.nih.gov/34000002/.

Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4:1256–1264, 2022. doi: 10.1038/s42256-022-00580-7. URL https://www.nature.com/articles/s42256-022-00580-7.

Yi-Ching Tang, Reid T. Powell, and Assaf Gottlieb. Molecular pathways enhance drug response prediction using transfer learning from cell lines to tumors and patient-derived xenografts. *Scientific Reports*, 12:16109, 2022. doi: 10.1038/s41598-022-20646-1. URL https://www.nature.com/articles/s41598-022-20646-1.

Zirui Wang, Zihang Dai, Barnabas Poczos, and Jaime Carbonell. Characterizing and avoiding negative transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11293–11302, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Wang_Characterizing_and_Avoiding_Negative_Transfer_CVPR_2019_paper.html.

Xiaoqiong Xia, Chaoyu Zhu, Yuqi Shan, Fan Zhong, Lei Liu, et al. Transcdr: a deep learning model for enhancing the generalizability of drug activity prediction through transfer learning and multimodal data fusion. *BMC Biology*, 22(1):227, 2024. doi: 10.1186/s12915-024-02023-8. URL https://pubmed.ncbi.nlm.nih.gov/39385185/.

Yingke Yang and Peiluan Li. Drug response prediction based on graph convolutional network and pathway enrichment analysis. *BMC Bioinformatics*, 24(1):471, 2023. doi: 10.1186/s12859-023-05592-1. URL https://pubmed.ncbi.nlm.nih.gov/38105227/.

Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. Provable dynamic fusion for low-quality multimodal data. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 41753–41769, 2023. URL https://proceedings.mlr.press/v202/zhang23ar.html.

Yuanpeng Zhang, Zhijian Huang, Ziyu Fan, Siyuan Shen, Yahan Li, Shangqian Wu, Min Wu, and Lei Deng. Deepsadr: Deep transfer learning with subsequence interaction and adaptive readout for cancer drug response prediction. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/bd4229b9f76ade256bb2f687e25e847d-Paper-Conference.pdf.

A Additional implementation details

Response node, hidden, latent, and gate-hidden dimensions are 64, dropout is .1, batch size is 32, learning rate is 5×10^{-4}, weight decay is 10^{-5}, and each response fit runs ten epochs with AdamW. The BCE objective includes KL weight .001. Drug features in the response branch are trainable fragment embeddings. Frozen evidence uses `ibm/MoLFormer-XL-both-10pct` and `facebook/esm2_t6_8M_UR50D`, followed by the trained pair encoder. Both OOD abstention and its learned-gate control use density normalization `none`. The SelectiveNet-style control uses a gate-coverage target .5 and penalty weight 32, with no hard density abstention configured. These are implementation-specific comparisons, not claims to reproduce every training choice of the cited methods.

The evidence support threshold uses the maximum binary pathway count among supported training drugs. In provenance mode, the count check occurs before soft confidence weighting; the 75th-percentile cap is calculated only on covered training pathways. Unsupported evidence has zero coverage. Current source-data matching and provenance scripts should be consulted before any cohort expansion: the original 338-row patient file includes 152 pseudonymous records whose TCGA cancer assignments cannot be verified, whereas the 186-row subset has GDC-resolvable identities. Neither cohort is the basis of the primary drug-shift tables.

Table 3 gives the exact learned-parameter count of each fusion module at the dimensions used in every saved run (node, hidden, and gate-hidden dimensions of 64), computed by instantiating each module directly rather than estimated from the layer equations. The gate shared by OOD abstention, the unnormalized-gate control, provenance weighting, and the SelectiveNet-style objective has roughly 25% more parameters than concatenation or cross-attention, so an improvement over those two controls is not attributable to a smaller model; an improvement over static addition, which has no fusion parameters at all, is a weaker comparison and is reported alongside the others rather than treated as capacity-matched.

Table 3: Learned-parameter count of each fusion module (node/hidden/gate-hidden dimension 64). Static addition performs no learned transform; the remaining figures come from direct module instantiation, not an estimate.

Fusion module	Parameters
Static addition	0
Concatenation	16,576
Cross-attention	16,768
Isolated-residual head	16,641
Gate (OOD abstention / unnormalized gate / provenance / SelectiveNet-style)	20,674

B Full reliability sensitivity summaries

All three-seed comparisons below use seeds 50–52; scaffold8 uses seeds 50–57 for zero and OOD abstention only. The conditional worst-utility intervals use 20,000 paired resamples of training seeds within each fixed fold, retaining the drugs in that fold together. LODO denotes deleting one drug from the already computed utility vector and taking the minimum of the remaining seven; no model is refit. These intervals do not include resampling cell lines, patients, new folds, or unseen drugs. Harm counts apply strict inequalities at exploratory absolute-AUC cutoffs. Table 5 reports the separate fold-resampling summaries for mean utility; neither table should be used as a population-level no-harm guarantee.

Table 4: Harm-count sensitivity and conditional uncertainty of worst utility. Each protocol has eight supported drugs.

Protocol	Method	H_0	$H_{.02}$	$H_{.05}$	Worst U interval	LODO range
heldout3	Adaptive readout (zero)	3/8	2/8	1/8	[-0.2518,-0.0165]	[-0.0864,-0.0247]
heldout3	Static addition	4/8	1/8	1/8	[-0.4463,-0.0074]	[-0.1588,-0.0105]
heldout3	Concatenation	5/8	1/8	1/8	[-0.3806,-0.0832]	[-0.2520,-0.0112]
heldout3	Cross-attention	5/8	2/8	0/8	[-0.1746,-0.0059]	[-0.0349,-0.0261]
heldout3	Learned gate	4/8	1/8	1/8	[-0.1628,-0.0024]	[-0.0756,-0.0016]
heldout3	OOD abstention	3/8	0/8	0/8	[-0.0037,-0.0009]	[-0.0016,-0.0011]
heldout3	Provenance	4/8	0/8	0/8	[-0.0222,-0.0016]	[-0.0043,-0.0021]
heldout3	SelectiveNet-style	4/8	1/8	1/8	[-0.3374,-0.0365]	[-0.1855,-0.0062]
scaffold3	Static addition	5/8	1/8	1/8	[-0.1426,-0.0087]	[-0.0837,-0.0073]
scaffold3	Concatenation	3/8	1/8	1/8	[-0.1977,-0.0725]	[-0.1296,-0.0061]
scaffold3	Cross-attention	6/8	1/8	0/8	[-0.0719,-0.0055]	[-0.0247,-0.0087]
scaffold3	Learned gate	3/8	1/8	1/8	[-0.3376,-0.0149]	[-0.2099,-0.0021]
scaffold3	OOD abstention	2/8	0/8	0/8	[-0.0037,-0.0013]	[-0.0021,-0.0001]
scaffold3	Provenance	3/8	0/8	0/8	[-0.0112,-0.0028]	[-0.0038,-0.0021]
scaffold8	OOD abstention	3/8	0/8	0/8	[-0.0271,-0.0028]	[-0.0111,-0.0049]

Table 5: Descriptive fold-bootstrap intervals for mean utility on seed-matched comparisons. Four folds are resampled; their training sets overlap.

Protocol	Method	Mean U	95% fold-bootstrap interval
heldout3	Adaptive readout (zero)	-0.0023	[-0.0246,+0.0171]
heldout3	Static addition	-0.0182	[-0.0635,+0.0102]
heldout3	Concatenation	-0.0294	[-0.0948,+0.0065]
heldout3	Cross-attention	-0.0051	[-0.0234,+0.0106]
heldout3	Learned gate	-0.0053	[-0.0253,+0.0102]
heldout3	OOD abstention	+0.0066	[-0.0007,+0.0140]
heldout3	Provenance	+0.0059	[-0.0011,+0.0130]
heldout3	SelectiveNet-style	-0.0189	[-0.0663,+0.0111]
scaffold3	Static addition	-0.0104	[-0.0307,+0.0027]
scaffold3	Concatenation	-0.0124	[-0.0436,+0.0061]
scaffold3	Cross-attention	-0.0045	[-0.0086,-0.0013]
scaffold3	Learned gate	-0.0224	[-0.0745,+0.0056]
scaffold3	OOD abstention	+0.0038	[+0.0017,+0.0060]
scaffold3	Provenance	+0.0006	[-0.0008,+0.0030]
scaffold8	OOD abstention	+0.0009	[-0.0011,+0.0036]

C PER-DRUG UTILITY

Table 6: OOD-abstention utility relative to an independently trained baseline on the same seeds.

Drug	Held-out (3 seeds)	Scaffold (3 seeds)	Scaffold (8 seeds)
Dasatinib	-0.0016	-0.0001	-0.0012
Erlotinib	+0.0084	+0.0093	+0.0094
Gefitinib	-0.0009	+0.0062	+0.0016
Pazopanib	+0.0065	+0.0098	+0.0116
Pemetrexed	+0.0212	+0.0048	+0.0003
Sorafenib	+0.0198	+0.0001	-0.0111
Sunitinib	-0.0011	-0.0021	-0.0049
Tamoxifen	+0.0006	+0.0027	+0.0012

Figure 2 and Section 4 report the applied-gate-versus-utility correlation using U_{train}, the comparison against an independently trained baseline. Table 7 repeats this per drug for held-out-drug OOD abstention alongside U_{off}, the fixed-checkpoint comparison defined in Section 3, computed from the same four-fold, three-seed corruption artifacts used in Section 4. The two utilities need not rank drugs the same way: sorafenib has $U_{\mathrm{off}} = 0$ by the same-checkpoint suppression identity, since its evidence density in this protocol always exceeds the training threshold and the applied gate is exactly

zero, while its U_{train} is $+0.0198$ relative to an independently trained model with different fitted parameters. Spearman correlation between mean applied α and U_{off} is 0.357 ($p = .385$, $n = 8$), compared with -0.143 ($p = .736$) against U_{train}; both remain statistically unresolved at this sample size, and the sign difference between them illustrates why the two reference computations should not be substituted for each other when interpreting gate behavior.

Table 7: Held-out-drug OOD abstention: mean applied gate α alongside the fixed-checkpoint (U_{off}) and independently trained (U_{train}) utility comparisons, four folds and three seeds.

Drug	Mean α	U_{off}	U_{train}
Dasatinib	0.1702	-0.0002	-0.0016
Erlotinib	0.4176	+0.0116	+0.0084
Gefitinib	0.0702	-0.0002	-0.0009
Pazopanib	0.3797	+0.0167	+0.0065
Pemetrexed	0.0767	+0.0054	+0.0212
Sorafenib	0.0000	+0.0000	+0.0198
Sunitinib	0.3041	-0.0016	-0.0011
Tamoxifen	0.0201	+0.0007	+0.0006

D FIXED-CHECKPOINT CORRUPTION

The following values reuse the four-fold, three-seed Tier-2 audit. Deltas use observed evidence in the same trained checkpoints. No-evidence training is a separate comparator and must not replace the observed reference in this table.

Table 8: Inference-only corruption: supported-drug macro-AUC and change from observed.

Condition	AUC	Delta
Observed	.5956	.0000
Zero / removal 100%	.5915	-.0041
Removal 25%	.5835	-.0120
Removal 50%	.5864	-.0092
Removal 75%	.5883	-.0072
Pathway shuffle 25%	.5955	-.0001
Pathway shuffle 50%	.5955	-.0000
Pathway shuffle 75%	.5954	-.0002
Pathway shuffle 100%	.5955	-.0000
Noise 25%	.5946	-.0010
Noise 50%	.5949	-.0007
Noise 75%	.5936	-.0019
Noise 100%	.5927	-.0029
Drug shuffle	.5919	-.0037